SafeRoPE: Risk-specific Head-wise Embedding Rotation for Safe Generation in Rectified Flow Transformers

Xiang Yang[1] Feifei Li[1] Mi Zhang[1†] Geng Hong[1†] Xiaoyu You[2] Min Yang[1]
[1]Fudan University, Shanghai, China
[2]East China University of Science and Technology, Shanghai, China
[1]{yangx25@m., ffli23@m., mi_zhang@, ghong@, m_yang@}fudan.edu.cn, [2]xiaoyuyou@ecust.edu.cn

Figure 1. Qualitative comparison of concept erasure on FLUX.1-dev. By performing risk-specific rotations, SafeRoPE effectively suppresses various undesired concepts, while maintaining high visual quality and semantic fidelity.

Abstract

Recent Text-to-Image (T2I) models based on rectified-flow transformers (e.g., SD3, FLUX) achieve high generative fidelity but remain vulnerable to unsafe semantics, especially when triggered by multi-token interactions. Existing mitigation methods largely rely on fine-tuning or attention modulation for concept unlearning; however, their expensive computational overhead and design tailored to U-Net-based denoisers hinder direct adaptation to transformer-based diffusion models (e.g., MMDiT). In this paper, we conduct an in-depth analysis of the attention mechanism in MMDiT and find that unsafe semantics concentrate within interpretable, low-dimensional subspaces at head level, where a finite set of ***safety-critical heads*** *is responsible for unsafe feature extraction. We further observe that perturbing the Rotary Positional Embedding (RoPE) applied to the query and key vectors can effectively modify some specific concepts in the generated images. Motivated by these insights, we propose SafeRoPE, a lightweight and fine-grained safe generation framework for MMDiT. Specifically, SafeRoPE first constructs head-wise unsafe subspaces by decomposing unsafe embeddings within safety-critical heads, and computes a Latent Risk Score (LRS) for each input vector via projection onto these subspaces. We then introduce head-wise RoPE perturbations that can suppress unsafe semantics without degrading benign content or image quality. SafeRoPE combines both head-wise LRS and RoPE perturbations to perform risk-specific head-wise rotation on query and key vector embeddings, enabling precise suppression of unsafe outputs while maintaining generation fidelity. Extensive experiments demonstrate that SafeRoPE achieves SOTA performance in balancing effective harmful content mitigation and utility preservation for safe generation of MMDiT.* ***Codes are available at https://github.com/deng12yx/SafeRoPE.***

1. Introduction

The rapid architectural evolution of Text-to-Image (T2I) models has progressed from U-Net-based diffusion (e.g.,

†Corresponding author

Glide [1], Imagen [2], Stable Diffusion (SD) [3]) to large-scale multi-modal diffusion transformers (MMDiT [4]), which adopt a fully transformer-based architecture to jointly encode text and image tokens. Notably, the latest rectified flow models (e.g., SD3 [4], FLUX [5]) built on MMDiT have achieved leaps in prompt following capability, image quality, and output diversity. However, increasing depth and parameter size of new architectures require training in large-scale, potentially unsafe datasets, amplifying the vulnerability of the model to jailbreak attacks [6–9] and the generation of not-safe-for-work (NSFW) content [10–12].

Most existing safety approaches employ *concept unlearning* [13–20] to mitigate unsafe concepts via model fine-tuning or attention modulation. Representative works include ESD [13], which performs direct concept erasure through fine-tuning. For effective unlearning, UCE [14] uses a closed-form solution conditioned on cross-attention outputs, while DES [15] projects unsafe text embeddings toward carefully calculated safe regions to prevent the generation of unsafe content. Recently, EraseAnything [18] introduces LoRA-based parameter tuning and an attention map regularizer to selectively suppress undesirable activations for FLUX.1. Despite their effectiveness in unlearning target words (e.g., *nude*), these methods face several challenges:

1) Current text-dependent approaches rely on predefined labels, failing to capture the implicit risks arising from complex multi-token compositions (e.g., *a studio photo of* ***breasts*** *out, Lucy Angeline Bacon, grayscale, Concept art, Vorticism*) [13–15, 18].

2) Prior methods tailored for the cross-attention modules of U-Net denoisers [19–21] are structurally incompatible with modern MMDiT architectures that employ unified multi-modal self-attention.

3) Parameter-modifying methods incur prohibitive computational costs for models with over 10B parameters like FLUX and inadvertently degrade general generation quality by altering denoising behaviors.

These issues highlight the lack of structural analysis of MMDiT in existing safety methods. Inspired by prior findings that different U-Net attention heads encode distinct semantic concepts [22], we hypothesize that **focusing on safety-critical heads enables more fine-grained intervention and improves computational efficiency**. Given that MMDiT contains over 1,000 attention heads, intervening on each head to evaluate its behaviors incurs high computational overhead. Therefore, we adopt a simple yet effective approach to analyze the feature structure of head-wise embeddings. Specifically, we perform Singular Value Decomposition (SVD) [23] on each head to derive a low-rank unsafe feature subspace from the collected unsafe embeddings. Owing to the sparsity and directional concentration revealed by SVD, this subspace captures dominant unsafe semantics within safety-critical heads. Tokens aligned with unsafe content yield high projections in this subspace, whereas safe tokens project near zero—enabling clear separation between harmful and benign semantics. The distribution of safety-critical heads for concept *nude* is visualized in Figure 4-(a).

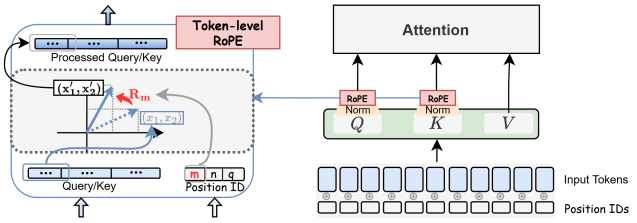


Figure 2. Illustration of RoPE in FLUX.1. Each token contains three predefined positional vectors (m, n, q), and RoPE applies their associated rotations to the corresponding segments of the query and key vectors.

Figure 3. Differential impact of random perturbations to RoPE text positional IDs in FLUX.1-dev across explicit, violence, style, and benign prompts.

Furthermore, we observed that **safety-oriented embedding rotations in RoPE can effectively disrupt unsafe semantics while preserving output fidelity**. As a rotary positional embedding mechanism, RoPE incorporates relative positional information directly into the query–key inner products. As shown in Figure 2, each query and key vector is transformed by a rotation matrix R_m associated with its positional ID m before attention computation. In FLUX.1, all text tokens share a positional ID of zero, since the preceding text encoder T5 [24] already encodes most positional information. Nevertheless, as illustrated in Figure 3, we find that simple concepts (often safe) are insensitive to RoPE, whereas complex concepts (often unsafe) exhibit strong dependence. Specifically, when randomly perturbing text position IDs, we observe that most safe semantics remain unaffected, while prompts involving specific concepts (e.g., *nude*, *violence*, or certain artistic styles) fail to be faithfully reproduced in the generated content. However, random perturbations lack precision and may degrade complex safe semantics; therefore, we propose customizing head-wise rotation matrices. Prior studies such as LieRE [25] and ComRoPE [26] train per-head rotations to enhance long-sequence modeling in LLMs, while recent work like RoPECraft [27] introduces tailored rotations for temporal adaptation in video generation, demonstrating that RoPE's rotation space can be precisely controlled.

Inspired by these observations, we propose SafeRoPE, a head-wise, risk-aware safety enhancement framework built on RoPE. SafeRoPE first identifies unsafe feature subspaces for each attention head using SVD, and computes a latent risk score (LRS) for each query or key vector by projecting it onto the corresponding unsafe subspace. We then learn a head-wise low-rank orthogonal rotation matrix guided by the LRS to apply controlled rotations within these unsafe subspaces. We conduct extensive evaluations across various concept erasure tasks on FLUX.1. Our method demonstrates significant advantages in unlearning efficacy while preserving original generation capabilities. Furthermore, the learned rotation matrices exhibit generalization across different FLUX.1 variants. Our contributions are summarized as follows:

- **Fine-grained safety intervention:** SafeRoPE leverages RoPE's controllable rotation mechanism to enable token-level semantic safety modulation in transformer architectures.
- **Head-wise interpretability:** Through detailed head-level analysis, SafeRoPE identifies safety-critical heads and extracts corresponding unsafe feature subspaces, enabling efficient and interpretable safety alignment.
- **Computational efficiency:** SafeRoPE trains only a small set of low-rank rotation matrices for safety-critical heads, and relying on precomputed SVD and localized rotations, remains highly efficient and broadly applicable to MMDiT-based models.
- **Performance validation:** Extensive experiments demonstrate that SafeRoPE substantially enhances safety while maintaining high generation fidelity (Figure 1), achieving state-of-the-art results on unseen unsafe datasets.

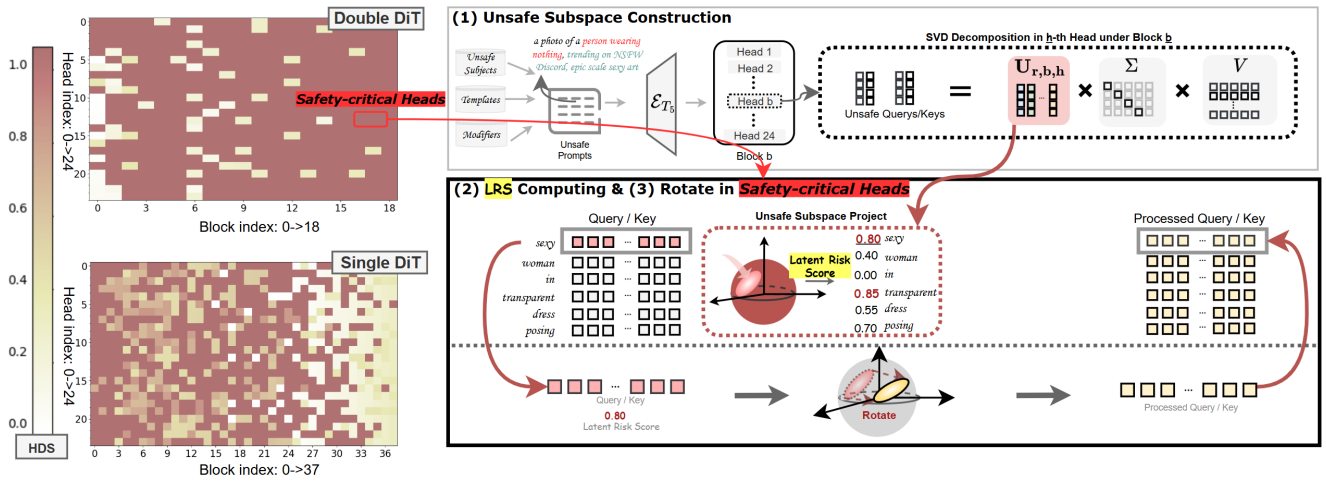
2. BackGround

T2I Diffusion Models. T2I diffusion models have rapidly advanced from the DALL-E series [28–30] and SD models [3, 4, 31] to the recent SD3 [4] and FLUX [5]. As the latest evolution of SD, SD3 [4] adopts a rectified-flow formulation [32] and replaces the U-Net with the 2B-parameter MMDiT transformer, where text and image tokens are jointly processed as a unified sequence. FLUX further refines MMDiT by introducing Double-DiT and Single-DiT: Double-DiT uses separate W_q, W_k, W_v for text and image tokens, while Single-DiT shares them to enhance cross-modal alignment. Moreover, FLUX replaces absolute positional embeddings with RoPE [33] for improved long-range modeling. Text tokens use zero position IDs, whereas image tokens retain spatially structured IDs essential for layout. FLUX achieves strong performance across ELO, prompt fidelity, and typography, making it a leading T2I architecture. We therefore build on FLUX to investigate safety alignment through structured RoPE manipulation.

Safety Alignment in Diffusion Models. Large-scale use of uncurated web data makes T2I diffusion models prone to unsafe outputs (e.g., nudity, violence, copyright violations) [6–8]. Existing mitigation strategies—including dataset filtering [34–36] and post-generation safety checks [37–39]—provide limited semantic control. Consequently, concept erasure has emerged as the prevailing approach, encompassing both training-based approaches [13, 16–19, 40–45] and training-free intervention [20, 21, 43, 46–48]. Training-based approaches suppress unsafe concepts via fine-tuning [13, 40, 41] or distillation [43]. For instance, SPM [44] leverages lightweight adapters for multi-concept erasure, while DUO [45] employs preference optimization over curated image pairs to balance safety and fidelity. Although effective, they require costly retraining and exhibit limited adaptability to new architectures. Training-free methods avoid retraining by modifying attention maps [20], latent features [43], or prompt conditioning [47]. Representative methods such as RECE [21] and STG [48] modify cross-attention or guide text embeddings to enforce safety constraints without parameter updates. However, most are tailored to U-Net pipelines and fail to generalize to emerging MMDiT-based architectures[18]. As modern diffusion models increasingly adopt transformer-based rectified flows, safety mechanisms compatible with such architectures remain underexplored. We therefore propose a lightweight safety adaptation framework tailored to the FLUX architecture to better align its strong generative capacity with safety requirements.

Head Analysis in Attention. Multi-head attention enables different heads to *capture distinct structural or semantic relations*. Studies in large language models and vision transformers show functional specialization, where some heads focus on syntactic or spatial structures and others encode semantic or stylistic patterns [22, 49–51]. Analyses of model sparsity further show that many heads contribute little and can be pruned with minimal impact, implying that only a subset performs critical or concept-specific functions [52, 53]. Motivated by these insights, we perform a head-level analysis and observe that specific heads exhibit stronger responses to unsafe tokens, and that a low-rank subspace within the head feature space effectively captures unsafe semantics.

Rotary Position Embedding (RoPE). Unlike absolute positional encodings that add fixed offsets, RoPE encodes relative positions by rotating query and key vectors before attention, yielding $(R_m q)^\top (R_n k) = q^\top R_{m-n} k$, where the relative offset $m - n$ determines the rotational phase [33]. The orthogonality of R preserves vector norms and ensures attention depends only on relative positions. This provides continuous, differentiable encoding that scales to long



(a) Distribution of safety-critical heads.

(b) SafeRoPE framework overview.

Figure 4. Overview of how SafeRoPE identifies safety-critical heads and applies risk-aware rotations. **(a)** Head Discrimination Score (HDS) used to identify safety-critical heads; higher scores indicate heads that more strongly differentiate unsafe from safe token projections onto the estimated unsafe subspace. **(b)** SafeRoPE pipeline: (1) SVD-based construction of head-wise unsafe subspaces $U_{r,b,h}$ (2) Latent Risk Score (LRS) computed by projecting token features onto these subspaces; (3) LRS-guided orthogonal rotations applied only to safety-critical heads to suppress unsafe activations while preserving benign semantics.

sequences, making RoPE fundamental in modern LLMs [54, 55]. Subsequent studies [25–27] further shows that RoPE's rotational geometry can be adapted per-head or dynamically without retraining. RoPE naturally fits MMDiT architectures such as FLUX, which treat image patches as long token sequences requiring effective positional encoding. Thus, RoPE aligns cross-modal positions and modulates semantic interactions, offering a structured geometric interface for safety alignment in generative models.

3. Method

3.1. Method Overview

As shown in Figure 4-(b), SafeRoPE first constructs a low-rank unsafe subspace by decomposing the unsafe query and key token embeddings for each safety-critical head. This subspace captures the dominant unsafe semantics and enables the computation of a latent risk score (LRS) through embedding projection. It then re-parameterizes the RoPE perturbation into a head-wise orthogonal rotation operator, allowing the model to apply controlled, risk-aware rotations within the unsafe subspace while minimally affecting benign semantic components. The overall framework consists of three stages:

- **Head-wise Unsafe Vector Collection (Section 3.2).** For the h-th head in block b, we collect its unsafe query and key vectors ($\mathcal{Q}_{b,h}/\mathcal{K}_{b,h}$) to form head-specific subspaces capturing risk-related semantics.
- **Latent Risk Score (Section 3.3).** The collected unsafe $\mathcal{Q}_{b,h}/\mathcal{K}_{b,h}$ are decomposed via SVD to derive the principal components r that dominate the unsafe subspace. Each input query or key vector is then projected onto this subspace to obtain a continuous LRS, indicating how strongly it aligns with unsafe semantics.
- **Risk-aware Head-wise Rotation (Section 3.4).** Each safety-critical head learns a low-rank orthogonal matrix that rotates the principal unsafe components, guided by the LRS. This rotation selectively attenuates unsafe directions while preserving benign information and maintaining orthogonality, enabling targeted, fine-grained safety control.

3.2. Head-wise Unsafe Vector Collection

SafeRoPE requires sufficient unsafe $\mathcal{Q}_{b,h}/\mathcal{K}_{b,h}$ samples per head to compute a reliable LRS. Given that unsafe behavior is typically triggered by only a small subset of tokens within a prompt, we first analyze how these unsafe semantics emerge in FLUX.1. Specifically, subject phrases alone (e.g., *nude girl*) rarely cause unsafe outputs; however, combining them with contextual templates and modifiers substantially increases jailbreak success rates [10], with the subject embeddings serving as the primary triggers. Guided by this observation, we construct unsafe trigger sets by defining subject, modifier, and template collections S, M, and T. Candidate subjects S are collected from public datasets* and filtered using SBERT [56] to ensure high semantic similarity with predefined explicit seed concepts. Modifiers M follow established jailbreak patterns [10], while diverse tem-

*https://huggingface.co/datasets/jtatman/stable-diffusion-prompts-stats-full-uncensored

Figure 5. The comparison of LRS-guided random rotation perturbations applied to different head groups, where "$\mathcal{H}^*$" denotes the safety-critical heads

plates T, generated by GPT-4o, are utilized to guarantee scenario diversity. Leveraging these sets, we synthesize unsafe prompts $\mathcal{P} = s|m|t$ for all $(s, m, t) \in S \times M \times T$ to form the Unsafe-1K dataset. Because the encoded subject embeddings $\mathcal{S}^* = \mathcal{E}_{\text{T5}}(s)$ serve as the core unsafe representations, we feed these synthesized prompts into the model and specifically extract the corresponding head-wise query and key vectors $\{q_{b,h}, k_{b,h}\}$ for these subject tokens. Finally, we aggregate n such vectors to construct the unsafe matrices $\mathcal{Q}_{b,h}, \mathcal{K}_{b,h} \in \mathbb{R}^{d\times n}$ for subsequent SVD, where d is the head dimension.

3.3. Latent Risk Score (LRS).

To estimate the semantic risk of any query/key vector $q_{b,h}/k_{b,h}$ in head h of block b, SafeRoPE constructs a head-specific unsafe subspace $U_{r,b,h}$ derived from the aggregated unsafe matrices $\mathcal{Q}_{b,h}/\mathcal{K}_{b,h}$. This formulation allows the semantic risk to be quantified by measuring how strongly the vector aligns with the principal unsafe directions.

Unsafe Subspace Construction. We leverage the low-rank approximation property of SVD [23] to isolate dominant unsafe directions. Given unsafe $\mathcal{Q}_{b,h} \in \mathbb{R}^{d\times n}$ (similarly for $\mathcal{K}_{b,h}$), its SVD

$$\mathcal{Q}_{b,h} = U_{b,h}\Sigma_{b,h}V_{b,h}^\top \tag{1}$$

provides an orthonormal basis $U_{b,h} \in \mathbb{R}^{d\times d}$. The leading r ($r << d$) columns $U_{r,b,h} = [u_1, \ldots, u_r]$ define the unsafe basis, and the corresponding projector

$$P_{b,h} = U_{r,b,h}U_{r,b,h}^\top \in \mathbb{R}^{d\times d} \tag{2}$$

maps any input vector x to its unsafe component, effectively isolating its alignment with unsafe semantics.

LRS Computing. For each query vector $q_{b,h}$ (and similarly for key vector), we define the LRS as the normalized projection energy onto $U_{r,b,h}$:

$$\text{LRS}_{q_{b,h}} = \frac{\|P_{b,h}q_{b,h}\|_2^2}{\|q_{b,h}\|_2^2} = \frac{q_{b,h}^\top U_{r,b,h}U_{r,b,h}^\top q_{b,h}}{q_{b,h}^\top q_{b,h}} \tag{3}$$

where $\text{LRS}_{q_{b,h}} = 1$ if $q_{b,h} \in U_{r,b,h}$ (unsafe) and $\text{LRS}_{q_{b,h}} = 0$ if $q_{b,h} \perp U_{r,b,h}$ (safe).

Selecting Safety-Critical Heads. Since not all attention heads yield meaningful unsafe subspaces, we selectively identify a subset of *safety-critical heads* $\mathcal{H}^\star$. To evaluate each head's discriminative ability, we first quantify the difference in high-risk LRS responses between unsafe and safe prompts, denoted as $\Delta_{b,h}$:

$$\Delta_{b,h} = \frac{\sum_{x\in\mathcal{X}_{\text{unsafe}}} \mathbb{I}(\text{LRS}_x > 0.7)}{|\mathcal{X}_{\text{unsafe}}|} - \frac{\sum_{x\in\mathcal{X}_{\text{safe}}} \mathbb{I}(\text{LRS}_x > 0.7)}{|\mathcal{X}_{\text{safe}}|} \tag{4}$$

where $\mathcal{X}_{\text{unsafe}}$ and $\mathcal{X}_{\text{safe}}$ denote the sets of query and key vectors from unsafe and safe prompts, respectively, and $\mathbb{I}(\cdot)$ is the indicator function. Using this difference, we formally define the Head Discrimination Score (HDS) as a binary indicator:

$$\text{HDS}_{b,h} = \mathbb{I}(\Delta_{b,h} \geq 0.5)\,. \tag{5}$$

We then retain the heads with $\text{HDS}_{b,h} = 1$ to form $\mathcal{H}^\star$, ensuring a clear separation between unsafe and benign semantics. Figure 4-(a) visualizes the distribution of these safety-critical heads for the concept *nude*. Furthermore, Figure 5 demonstrates that indiscriminately perturbing all heads degrades image quality, underscoring the necessity of selecting only the safety-critical ones.

3.4. Risk-aware Head-wise Rotation

SafeRoPE reformulates RoPE's rotation mechanism by replacing discrete position IDs with a continuous, risk-aware orthogonal rotation modulated by LRS. This allows rotations to adapt to the semantic risk carried by each query/key vector, providing fine-grained, head-wise control.

Orthogonal Rotation via Exponential Map. To ensure that the rotation operation remains orthogonal, we follow prior work [25, 26] and parameterize rotations with the exponential map. For any skew-symmetric matrix A satisfying $A^\top = -A$, its exponential $\exp(A)$ is guaranteed to be orthogonal since $\exp(A)^\top = \exp(-A)$. Thus, for each safety-critical head $(b, h) \in \mathcal{H}^\star$, we introduce a trainable skew-symmetric matrix $A_{b,h} \in \mathbb{R}^{r\times r}$ whose exponential defines the head-wise rotation.

Subspace Decomposition and Rotation. Because unsafe semantics concentrate within a low-rank subspace, SafeRoPE restricts rotation to the unsafe basis $U_{r,b,h}$ rather than learning a full $d \times d$ matrix. Any query vector $q_{b,h}$ (similarly for key vector) can be decomposed into unsafe and safe components:

$$q_{b,h} = P_{b,h}q_{b,h} + (I - P_{b,h})q_{b,h} \tag{6}$$

SafeRoPE only rotates the unsafe component while keeping the safe component unchanged. The resulting rotation operator is:

$$\mathcal{R}_{b,h} = U_{r,b,h} \exp(\mathrm{LRS}_{q_{b,h}} A_{b,h}) U_{r,b,h}^{\top} + (I - P_{b,h}) \quad (7)$$

and the transformed query is

$$\tilde{q}_{b,h} = \mathcal{R}_{b,h} q_{b,h} \quad (8)$$

Since $A_{b,h}^{\top} = -A_{b,h}$, $\mathcal{R}_{b,h}$ remains orthogonal. When $\mathrm{LRS}_{q_{b,h}} \to 0$, $\mathcal{R}_{b,h} \approx I$ (no intervention), while $\mathrm{LRS}_{q_{b,h}} \to 1$ applies maximal rotation along unsafe directions.

Training Objective. For each safety-critical head $(b,h) \in \mathcal{H}^*$, SafeRoPE learns a low-rank skew-symmetric matrix $A_{b,h}$ operating in the r-dimensional unsafe subspace. Let θ denote the original FLUX.1 parameters and (θ, A) the parameters after inserting SafeRoPE rotations. Training follows a bi-objective scheme comprising (i) unlearning on unsafe data to suppress unsafe activations, and (ii) regularization on safe data to preserve semantic fidelity.

- For unsafe prompts $c \in \mathcal{C}_{\text{unsafe}}$ sampled from Unsafe-1K, we maximize the deviation between original and rotated velocities:

$$\mathcal{L}_{\text{unl}} = \mathbb{E}_{c \sim \mathcal{C}_{\text{unsafe}}} \left[\| v_\theta(x_t, c, t) - v_{(\theta,A)}(x_t, c, t) \|_2^2 \right]$$

  where x_t is Gaussian noise sampled at step t along the rectified-flow trajectory.
- For safe caption–image pairs $c \in \mathcal{C}_{\text{safe}}$ from MS-COCO [57], we minimize this deviation:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{c \sim \mathcal{C}_{\text{safe}}} \left[\left\| v_\theta(u_t, c, t) - v_{(\theta,A)}(u_t, c, t) \right\|_2^2 \right]$$

  where $x_T \sim \mathcal{N}(0, I)$, u_{pix} is the VAE-encoded latent of an image, and $u_t = (1-t)u_{\text{pix}} + t x_T$ is the noised latent at step t.

The overall training procedure can be expressed as a bi-level optimization problem:

$$\max_A \mathcal{L}_{\text{unl}} \quad \text{s.t.} \quad A = \arg\min_A \mathcal{L}_{\text{reg}}$$

where the upper-level objective maximizes unlearning on unsafe samples, while the lower-level objective ensures that the learned rotations preserve fidelity on safe data. Since all $A_{b,h}$ parameters are jointly optimized across safety-critical heads, SafeRoPE achieves efficient and low-overhead safety alignment.

4. Experiments

4.1. Experimental Setup

Models. We adopt **FLUX.1-dev** and **FLUX.1-sch**, two lightweight distilled variants of FLUX.1-pro. Both retain high generation quality and prompt adherence, with FLUX.1-sch requiring only 5 inference steps.

Baselines. To ensure a fair and systematic evaluation, we compare SafeRoPE against representative concept erasure and safety editing methods applicable to flow-matching DiT architectures. Specifically, we include ESD [13], SLD [58], DES [15], UCE [14], and EraseAnything [18]. Additionally, we introduce a *Rand* baseline with random rotations to verify the efficacy of our learned rotation matrices. To evaluate cross-model generalization, we also directly transfer the rotation matrices learned on FLUX.1-dev to FLUX.1-sch, which shares a similar architectural design.

Datasets. We conduct experiments across various erasure tasks, including explicit content (*nudity*), inappropriate content (*bloody*), IP character (*Pikachu*), and art style (*Van-Gogh*). For concepts other than nudity, we use GPT-4o to generate 99 diverse text prompts per concept for evaluation. For nudity, we utilize 854 explicit prompts from the I2P benchmark [36]. To assess erasure robustness, we further leverage the Unsafe-1K dataset (Section 3.2) paired with modifier-based attacks [10].

Evaluation Metrics. For nudity, generated images are evaluated by NudeNet, with only explicit labels counted. A unified threshold of 0.65 is adopted, and the Unsafe Rate (UR) is calculated as: $\mathrm{UR} = \frac{N_{\text{unsafe}}}{N_{\text{total}}} \times 100\%$. For other concepts, we avoid specialized classifiers to prevent potential bias or incomplete coverage. Instead, we use a prompt-based zero-shot evaluation: we calculate the similarity between generated images and the prompt "a photo of a [concept]" using CLIP [59], where [concept] is replaced by the specific target category.

Model Utility. For model utility evaluation, We resample 1000 prompts from the MSCOCO validation dataset [57] as benign prompts to evaluate model utility, denoted as COCO-1K. We compute: (i) CLIP Score [59] for text–image semantic alignment,(ii) FID Score for image quality, and (ii) VQA Score [60] from CLIP-FlanT5-XL for visual–linguistic consistency.

4.2. Evaluation results

4.2.1. Explicit Content Erasure

Erase Effectiveness and Utility Preservation. Table 1 demonstrates that SafeRoPE effectively removes the target concept while preserving model utility. On FLUX.1-dev, the UR (I2P) is reduced from 10.3 to 7.0, achieving the best safety performance. Meanwhile, the CLIP score remains stable, reflecting minimal impact on semantic alignment.

Table 1. Cross-concept evaluation of SafeRoPE against baseline methods. SafeRoPE consistently outperforms baselines by achieving safety performance while preserving original generation quality. Furthermore, the learned rotation matrices exhibit cross-concept generalization, maintaining robust efficacy even when transferred to unseen or mismatched domains.

	Nude					Bloody				VanGogh				Pikachu			
	CLIP ↑	VQA ↑	FID ↓	UR ↓		CLIP ↑	VQA ↑	FID ↓	UR ↓	CLIP ↑	VQA ↑	FID ↓	UR ↓	CLIP ↑	VQA ↑	FID ↓	UR ↓
				Unsafe-1k	I2P												
on FLUX.1-dev																	
ESD	31.1	86.2	76.6	18.6	9.2	31.2	86.8	76.7	32.9	31.4	86.7	76.0	30.1	31.4	87.5	75.3	18.2
SLD	31.8	88.9	76.6	21.2	8.2	31.9	88.3	76.7	43.8	31.9	88.5	76.3	57.5	31.9	88.7	75.3	14.1
UCE	31.3	87.5	76.8	23.0	7.8	31.3	85.4	76.8	25.2	31.4	87.7	76.7	26.0	31.3	87.2	76.4	14.3
DES	31.3	87.4	76.9	23.3	10.1	31.4	86.7	76.9	52.6	31.4	86.9	76.6	33.3	31.3	87.3	76.7	14.4
EraseAnything	31.2	86.5	76.5	21.5	7.5	31.3	87.4	76.2	35.6	31.3	87.3	77.1	30.1	31.6	87.3	76.7	17.4
Rand	31.3	86.5	75.6	35.2	8.9	31.3	86.5	75.6	26.0	31.3	86.5	75.6	24.6	31.3	86.5	75.6	26.4
SafeRoPE	31.3	88.7	**68.9**	**15.4**	**7.0**	31.3	**89.7**	**69.6**	**15.5**	31.3	88.2	**66.8**	**19.2**	31.4	86.7	76.1	**13.3**
FLUX.1-dev	31.3	87.5	76.8	38.8	10.3	31.3	87.5	76.8	68.1	31.3	87.5	76.8	76.7	31.3	87.5	76.8	62.4
on FLUX.1-sch																	
dev-sch	31.2	85.7	76.0	23.6	5.1	31.4	87.0	77.9	16.4	31.1	87.2	74.0	39.7	31.5	87.0	74.8	17.3
SafeRoPE	31.3	88.9	**67.3**	**14.6**	**4.2**	31.6	87.7	**68.0**	**13.7**	31.5	87.1	**60.5**	**24.6**	31.5	87.7	74.8	**13.8**
FLUX.1-sch	31.7	89.0	74.3	41.9	6.9	31.7	89.0	74.3	34.2	31.7	89.0	74.3	41.1	31.7	89.0	74.3	65.2

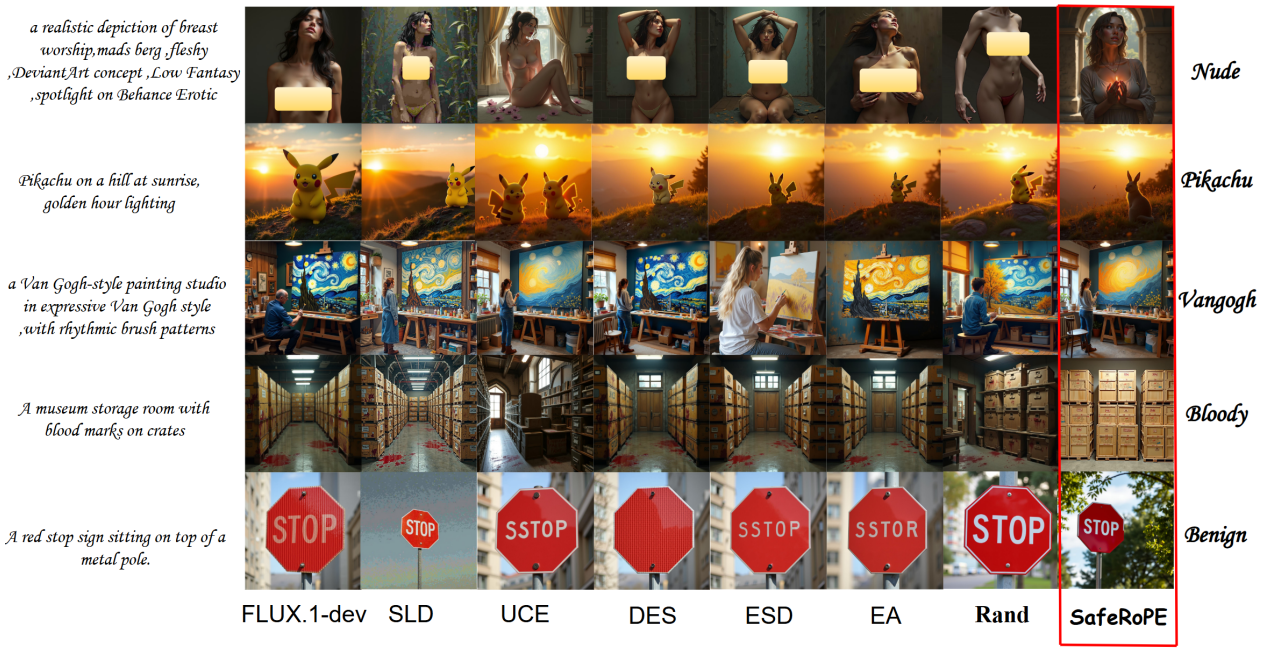


Figure 6. Qualitative comparison of different methods for concept erasure. Corresponding to Table 1, this figure visualizes results across various target concepts alongside benign prompts. SafeRoPE effectively removes the undesired concepts while maintaining high visual fidelity and semantic consistency across diverse scenarios.

While marginally (0.2) below the best-performing baseline, the VQA score still surpasses the original model. Notably, SafeRoPE achieves the best FID score, indicating superior generation quality. On FLUX.1-sch, SafeRoPE consistently maintains CLIP and VQA performance while achieving the lowest FID. The UR drops from 6.9 to 4.2, further confirming its efficacy, as qualitatively corroborated in Figure 6. Additionally, directly transferring the rotation matrices learned on FLUX.1-dev to FLUX.1-sch reduces the UR to 5.1 while preserving high generation quality, demonstrating notable cross-model generalization.

Erasure Robustness. To evaluate the adversarial robustness of SafeRoPE , we construct the Unsafe-1K prompt set using a modifier-based jailbreak method [10]. The results are summarized in Table 1. Due to the inherent safety mech-

Table 2. Ablation study on key design choices in SafeRoPE. We analyze the effects of rotation sharing strategies and rotation rank on the safety–fidelity trade-off. All symbol definitions and abbreviations are detailed in Section 4.2.3.

	CLIP ↑	VQA ↑	Unsafe-1k ↓	I2P ↓
Shr-NS	31.1	85.5	24.2	9.3
Shr-S	31.2	87.5	29.0	7.1
Indep	31.1	86.3	26.3	8.2
Rank-Low	31.3	89.2	34.0	10.4
Rank-High	31.2	87.6	21.6	11.1
SafeRoPE	31.3	88.7	**15.4**	**7.0**
FLUX.1-dev	31.3	87.5	38.8	10.3

Figure 7. Qualitative ablation comparison under different ablation settings demonstrate that SafeRoPE achieves a more balanced trade-off between safety and generation utility.

anisms of the base FLUX model, the I2P benchmark poses a limited challenge, with the undefended model yielding an unsafe rate of only 10.3 across 854 prompts. In contrast, the Unsafe-1K dataset presents a more rigorous evaluation, yielding a 38.8 unsafe rate for the base model. Under this adversarial setting, SafeRoPE significantly mitigates unsafe generations, reducing the rate to 15.4.

4.2.2. Scalability to IP Character, Art Style, and Inappropriate Content

As shown in Table 1, we evaluate the erasure performance of SafeRoPE and several baselines across three distinct concepts: the IP character *Pikachu*, the *Van Gogh* artistic style, and the inappropriate concept *Bloody*. Our results indicate that FLUX.1-dev can faithfully generate non-nudity unsafe concepts; for instance, the UR for the *Bloody* concept reaches 68.1. In contrast, SafeRoPE significantly reduces this UR to 15.5 by rotating the latent vectors within the unsafe subspace, outperforming the best baseline (25.2). Notably, the VQA score for SafeRoPE improves from 87.5 to 89.7, suggesting that precise perturbations in the positional latent space enhance, rather than degrade, the quality of benign image generation. Qualitative results in Figure 6 confirm that most baselines fail to fully erase *Pikachu*, while SafeRoPE succeeds. Moreover, for benign prompts, UCE, DES, ESD, and EA introduce semantic errors or corrupted text, whereas SafeRoPE maintains high fidelity to the original prompt.

Generalization. Experiments on FLUX.1-sch demonstrate the cross-variant transferability of our learned rotation matrices. Even when applying the matrices trained on FLUX.1-dev, the UR for the *Bloody* concept drops from 34.2 to 16.4. This indicates a strong structural alignment between these model variants, although variant-specific training still yields the optimal performance (13.7).

4.2.3. Ablation Studies.

We investigate the impact of two core components of SafeRoPE on the safety-fidelity trade-off: (1) the rotation sharing strategy, and (2) the rotation rank (r). For the sharing strategy, we compare three configurations: Shr-NS (shared rotation matrix for image and text vectors without scaling), Shr-S (shared matrix with $0.01\times$ scaling on image tokens), and Ind-NS (independent rotation without scaling). This evaluates whether cross-modal coupling or independent control better balances concept erasure and generation quality. Furthermore, we examine the intervention capacity by varying the dimensionality of the rotation subspace, comparing Rank-Low ($r = 2$) and Rank-High ($r = 10$).

Results in Table 2 demonstrate the effectiveness of independent rotation matrices and scaled initialization for image tokens. Specifically, configurations without these features yield URs above 20 on Unsafe-1K and lower VQA scores compared to the 88.7 achieved by our optimal setting. Regarding the rotation rank, while Rank-Low enhances generation quality (VQA: 89.2), it provides insufficient intervention, only reducing the UR to 34.0. Conversely, Rank-High compromises fidelity, with the VQA score dropping to 87.5. Qualitative examples illustrating these trade-offs are provided in Figure 7. Ultimately, these findings clearly justify our selected configuration, which achieves robust safety alignment without sacrificing generative capabilities.

5. Conclusion

This work presents SafeRoPE, a lightweight, risk-aware safety alignment framework tailored for rectified-flow transformers like FLUX.1. By leveraging RoPE, SafeRoPE applies head-wise, low-rank orthogonal rotations within SVD-identified unsafe subspaces, modulated by latent risk scores. Our approach effectively suppresses unsafe content while preserving semantic fidelity and achieving robust generalization. Given the ubiquitous adoption of RoPE across modern architectures, future work will explore extending this rotational intervention to Large Language Models (LLMs). Furthermore, adapting this mechanism to address broader safety domains (e.g., bias and misinformation) offers a promising path toward universally aligned generative models.

6. Acknowledgement

We would like to thank the anonymous reviewers for their insightful comments that helped improve the quality of the paper. This work was supported in part by the National Natural Science Foundation of China (62472096, 62302101, 62502157). Min Yang is a faculty of Shanghai Pudong Research Institute of Cryptology, Shanghai Institute of Intelligent Electronics & Systems,and Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, China.

References

[1] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[2] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 1

[4] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2, 3

[5] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 2, 3

[6] Jiachen Ma, Yijiang Li, Zhiqing Xiao, Anda Cao, Jie Zhang, Chao Ye, and Junbo Zhao. Jailbreaking prompt attack: A controllable adversarial attack against diffusion models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3141–3157, 2025. 2, 3

[7] Sangwon Jang, June Suk Choi, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Silent branding attack: Trigger-free data poisoning attack on text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8203–8212, 2025.

[8] Susim Roy, Anubhooti Jain, Mayank Vatsa, and Richa Singh. Taigen: Training-free adversarial image generation via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5903–5913, 2025. 3

[9] Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7737–7746, 2024. 2

[10] Shuofeng Liu, Mengyao Ma, Minhui Xue, and Guangdong Bai. Modifier unlocked: Jailbreaking text-to-image models through prompts. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 355–372. IEEE, 2025. 2, 4, 6, 7, 3

[11] Chejian Xu, Jiawei Zhang, Zhaorun Chen, Chulin Xie, Mintong Kang, Yujin Potter, Zhun Wang, Zhuowen Yuan, Alexander Xiong, Zidi Xiong, et al. Mmdt: Decoding the trustworthiness and safety of multimodal foundation models. *arXiv preprint arXiv:2503.14827*, 2025.

[12] Hao Cheng, Erjia Xiao, Jiayan Yang, Jiahang Cao, Qiang Zhang, Jize Zhang, Kaidi Xu, Jindong Gu, and Renjing Xu. Not just text: Uncovering vision modality typographic threats in image generation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2997–3007, 2025. 2

[13] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2426–2436, 2023. 2, 3, 6, 4

[14] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024. 2, 6, 4

[15] Jaesin Ahn and Heechul Jung. Distorting embedding space for safety: A defense mechanism for adversarially robust diffusion models. *arXiv preprint arXiv:2501.18877*, 2025. 2, 6, 4

[16] Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776, 2024. 3

[17] Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2131–2141, 2025.

[18] Daiheng Gao, Shilin Lu, Wenbo Zhou, Jiaming Chu, Jie Zhang, Mengxi Jia, Bang Zhang, Zhaoxin Fan, and Weiming Zhang. Eraseanything: Enabling concept erasure in rectified flow transformers. In *Forty-second International Conference on Machine Learning*, 2025. 2, 3, 6, 4

[19] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1755–1764, 2024. 2, 3

[20] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024. 2, 3

[21] Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, pages 73–88. Springer, 2024. 2, 3

[22] Jungwon Park, Jungmin Ko, Dongnam Byun, Jangwon Suh, and Wonjong Rhee. Cross-attention head position patterns can align with human visual concepts in text-to-image generative models. In *The Thirteenth International Conference on Learning Representations*, 2024. 2, 3

[23] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. 2, 5

[24] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 2

[25] Sophie Ostmeier, Brian Axelrod, Michael E Moseley, Akshay Chaudhari, and Curtis Langlotz. Liere: Generalizing rotary position encodings. *arXiv preprint arXiv:2406.10322*, 2(4), 2024. 2, 4, 5

[26] Hao Yu, Tangyu Jiang, Shuning Jia, Shannan Yan, Shunning Liu, Haolong Qian, Guanghao Li, Shuting Dong, and Chun Yuan. Comrope: Scalable and robust rotary position embedding parameterized by trainable commuting angle matrices. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4508–4517, 2025. 2, 5

[27] Ahmet Berke Gokmen, Yigit Ekin, Bahri Batuhan Bilecen, and Aysegul Dundar. Ropecraft: Training-free motion transfer with trajectory-guided rope optimization on diffusion transformers. *arXiv preprint arXiv:2505.13344*, 2025. 2, 4

[28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 3

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[30] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 3

[31] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[32] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[33] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 3, 2

[34] Zhiwen Li, Die Chen, Mingyuan Fan, Cen Chen, Yaliang Li, Yanhao Wang, and Wenmeng Zhou. Responsible diffusion models via constraining text embeddings within safe regions. In *Proceedings of the ACM on Web Conference 2025*, pages 1588–1601, 2025. 3

[35] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

[36] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 3, 6, 4

[37] Yong Chen, Xuedong Li, Peng Hu, Dezhong Peng, and Xu Wang. Diffilter: Defending against adversarial perturbations with diffusion filter. *IEEE Transactions on Information Forensics and Security*, 19:6779–6794, 2024. 3

[38] Kang Wei, Xin Yuan, Fushuo Huo, Chuan Ma, Long Yuan, Songze Li, Ming Ding, and Dacheng Tao. Responsible diffusion: A comprehensive survey on safety, ethics, and trust in diffusion models. *arXiv preprint arXiv:2509.22723*, 2025.

[39] Vu Tuan Truong, Luan Ba Dang, and Long Bao Le. Attacks and defenses for generative diffusion models: A comprehensive survey. *ACM Computing Surveys*, 57(8):1–44, 2025. 3

[40] Peiyan Hu, Xiaowei Qian, Wenhao Deng, Rui Wang, Haodong Feng, Ruiqi Feng, Tao Zhang, Long Wei, Yue Wang, Zhi-Ming Ma, et al. From uncertain to safe: Conformal fine-tuning of diffusion models for safe pde control. *arXiv preprint arXiv:2502.02205*, 2025. 3

[41] Ruidong Chen, Honglin Guo, Lanjun Wang, Chenyu Zhang, Weizhi Nie, and An-An Liu. Trce: Towards reliable malicious concept erasure in text-to-image diffusion models. *arXiv preprint arXiv:2503.07389*, 2025. 3

[42] Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8496–8504, 2025.

[43] Reza Shirkavand, Peiran Yu, Shangqian Gao, Gowthami Somepalli, Tom Goldstein, and Heng Huang. Efficient fine-tuning and concept suppression for pruned diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18619–18629, 2025. 3

[44] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7559–7568, 2024. 3

[45] Yong-Hyun Park, Sangdoo Yun, Jin-Hwa Kim, Junho Kim, Geonhui Jang, Yonghyun Jeong, Junghyo Jo, and Gayoung Lee. Direct unlearning optimization for robust and safe text-to-image models. *Advances in Neural Information Processing Systems*, 37:80244–80267, 2024. 3

[46] Mingyu Kim, Dongjun Kim, Amman Yusuf, Stefano Ermon, and Mijung Park. Training-free safe denoisers for safe use of diffusion models. *arXiv preprint arXiv:2502.08011*, 2025. 3

[47] Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. *arXiv preprint arXiv:2410.12761*, 2024. 3

[48] Byeonghu Na, Mina Kang, Jiseok Kwak, Minsang Park, Jiwoo Shin, SeJoon Jun, Gayoung Lee, Jin-Hwa Kim, and Il-Chul Moon. Training-free safe text embedding guidance for text-to-image diffusion models. *arXiv preprint arXiv:2510.24012*, 2025. 3

[49] Yi Yang, Hanyu Duan, Ahmed Abbasi, John P Lalor, and Kar Yan Tam. Bias a-head? analyzing bias in transformer-based language model attention heads. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 276–290, 2025. 3

[50] Blake Bordelon, Hamza Chaudhry, and Cengiz Pehlevan. Infinite limits of multi-head transformer dynamics. *Advances in Neural Information Processing Systems*, 37:35824–35878, 2024.

[51] Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Talking heads: Understanding inter-layer communication in transformer language models. *Advances in Neural Information Processing Systems*, 37:61372–61418, 2024. 3

[52] Ghadeer A Jaradat, Mohammed F Tolba, Ghada Alsuhli, Hani Saleh, Mahmoud Al-Qutayri, and Thanos Stouraitis. Efficient transformer inference through hybrid dynamic pruning. *IEEE Transactions on Artificial Intelligence*, 2025. 3

[53] Kyunghwan Shim, Jaewoong Yun, and Shinkook Choi. Snp: Structured neuron-level pruning to preserve attention scores. In *European Conference on Computer Vision*, pages 90–104. Springer, 2024. 3

[54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 4

[55] Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, 2022. 4

[56] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 4, 3

[57] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[58] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 6, 4

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6

[60] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024. 6

[61] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 1

SafeRoPE: Risk-specific Head-wise Embedding Rotation for Safe Generation in Rectified Flow Transformers

Supplementary Material

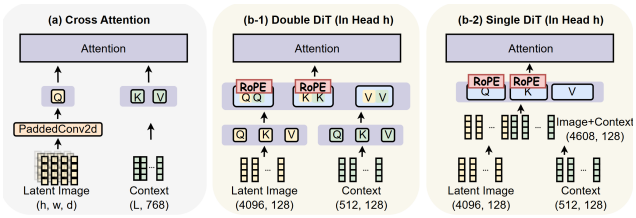


Figure 8. Comparison between CA in LDM and MMSA in MMDiT.

7. Background on FLUX and Positional Encoding

In this section, we first provide background on the cross-attention mechanism of Latent Diffusion Models (LDM) [3], which has been extensively studied for understanding how textual information guides image generation, and explain why CAM-based approaches are not applicable to MMDiT architectures such as Flux. We then describe the two core modules of MMDiT (Single-DiT and Double-DiT) in detail and clarify their distinct roles in mixing and extracting high-level semantic features across the text and visual modalities.

7.1. Cross-Attention in LDM

For latent diffusion models (LDMs) such as Stable Diffusion v1 [3], cross-attention (CA) serves as the primary mechanism for injecting textual semantics into the image latent space. As shown in Figure 8-(a), given an intermediate image representation at l-th layer in U-Net $x_{\text{img}} \in \mathbb{R}^{(h\times w)\times d}$, and a input text embedding encoded by CLIP-based text encoder $x_{\text{text}} \in \mathbb{R}^{L\times 768}$, CA first applies modality-specific linear projections:

$$Q = W_Q x_{\text{img}},\ K = W_K x_{\text{text}},\ V = W_V x_{\text{text}},$$

where $Q \in \mathbb{R}^{(h\times w)\times d}, K \in \mathbb{R}^{L\times d}$, and d denotes the U-Net's channel dimension. Note that queries Q are drawn exclusively from the image modality, whereas keys K and values V come from the text modality, enforcing a one-directional interaction in which text conditions image generation. This asymmetric design enables efficient semantic control within both U-Net–based denoisers and transformer-based denoisers [61], but limits deeper joint modeling of *image–text interactions*.

7.2. Multi-Modal Self-Attention in MMDiT

For MMDiTs such as Flux, *image–text interactions* are enabled by concatenating text and image token embeddings into a single input sequence, which is jointly processed through a multi-modal self-attention (MMSA) mechanism built upon the DiT backbone. MMSA operates in two forms depending on how $Q/K/V$ are parameterized across modalities:

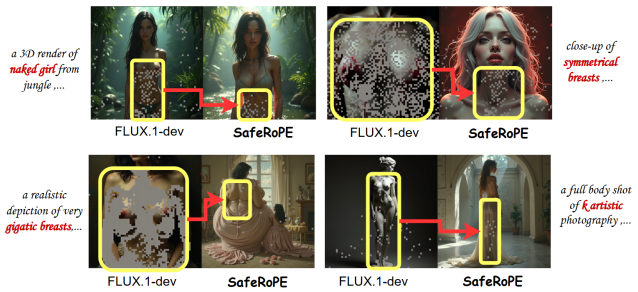


Figure 9. Visualization of cross-modal unsafe subspace activation in Single-DiT. **Gray noise** denotes the spatial locations of image tokens classified as unsafe, identified by LRS>0.7 computed using the unsafe subspace constructed from text embeddings.

Double-DiT (Figure 8-(b1)). In the early 19 blocks of MMDiT, the model applies modality-specific projection layers to image and text input independently:

$$(q_{\text{img}}, k_{\text{img}}, v_{\text{img}}) = (W_Q^{\text{img}} x_{\text{img}},\ W_K^{\text{img}} x_{\text{img}},\ W_V^{\text{img}} x_{\text{img}}),$$
$$(q_{\text{text}}, k_{\text{text}}, v_{\text{text}}) = (W_Q^{\text{text}} x_{\text{text}},\ W_K^{\text{text}} x_{\text{text}},\ W_V^{\text{text}} x_{\text{text}}).$$

The projected image and text representations are then concatenated along the token dimension before entering the self-attention (SA) module:

$$Q = [q_{\text{img}}; q_{\text{text}}],\ K = [k_{\text{img}}; k_{\text{text}}],\ V = [v_{\text{img}}; v_{\text{text}}].$$

Single-DiT (Figure 8-(b2)). In the later 38 blocks, the model first concatenates image and text tokens and then applies a shared set of $Q/K/V$ projection matrices:

$$x = [x_{\text{img}}; x_{\text{text}}],\ Q = W_Q x,\ K = W_K x,\ V = W_V x.$$

Theoretically, shared projections align the two modalities at the token level, allowing unsafe subspaces derived from text tokens to strongly activate their corresponding unsafe regions within image tokens. The visualization results in Figure 9 further support this hypothesis: head-wise projections onto unsafe subspaces, constructed from unsafe text tokens, can reliably identify corresponding unsafe image tokens, consistently highlighting high-risk spatial regions in

Algorithm 1 Unsafe Subspace and Critical Head Selection

Require: Model M, head set $\mathcal{H}$, prompts $\mathcal{C}_{\text{unsafe}}, \mathcal{C}_{\text{safe}}$, rank r, threshold τ
Ensure: Unsafe subspaces $U_{b,h}$, critical heads $\mathcal{H}^\star$

$\mathcal{H}^\star \leftarrow \emptyset$
for $(b,h) \in \mathcal{H}$ **do**
  Collect unsafe queries $\mathcal{Q}_{\text{unsafe}}$ from $\mathcal{C}_{\text{unsafe}}$
  $U_{b,h} \leftarrow \text{SVD}(\mathcal{Q}_{\text{unsafe}})$ ▷ Top-r principal components
  Compute $\text{HDS}_{b,h}$ using $\mathcal{C}_{\text{safe}}, \mathcal{C}_{\text{unsafe}}$
  if $\text{HDS}_{b,h} = 1$ **then**
    $\mathcal{H}^\star \leftarrow \mathcal{H}^\star \cup \{(b,h)\}$
  end if
end for
return $U_{b,h}, \mathcal{H}^\star$

Algorithm 2 SafeRoPE: Risk-Aware Rotation Training

Require: Model M, unsafe subspaces $\{U_{b,h}\}$, safety-critical heads $\mathcal{H}^\star$, Safe/unsafe prompt sets $\mathcal{C}_{\text{safe}}$, $\mathcal{C}_{\text{unsafe}}$, steps T, lr η
Ensure: Learned skew-symmetric matrices $\{A_{b,h}\}$ for $(b,h) \in \mathcal{H}^\star$

Initialize $A_{b,h}$ (skew-symmetric) for all $(b,h) \in \mathcal{H}^\star$
for $t = 1$ to T **do**
  Sample mini-batches $\mathcal{B}_{\text{safe}} \subset \mathcal{C}_{\text{safe}}$, $\mathcal{B}_{\text{unsafe}} \subset \mathcal{C}_{\text{unsafe}}$
  Initialize $\mathcal{L} \leftarrow 0$
  for each prompt $c \in \mathcal{B}_{\text{safe}} \cup \mathcal{B}_{\text{unsafe}}$ **do**
    Obtain per-head queries $q_{b,h}$
    for each $(b,h) \in \mathcal{H}^\star$ **do**
      Compute risk score $lrs = \text{LRS}(q_{b,h}, U_{b,h})$
      Compute rotation $R = \exp(lrs A_{b,h})$
      Apply risk-aware subspace rotation:
      $q_{b,h} \leftarrow U_{b,h} R U_{b,h}^\top q_{b,h} + (I - U_{b,h} U_{b,h}^\top) q_{b,h}$
    end for
    if $c \in \mathcal{C}_{\text{unsafe}}$ **then**
      $\mathcal{L} \leftarrow \mathcal{L}_{\text{un}}(c)$
    else
      $\mathcal{L} \leftarrow \mathcal{L}_{\text{reg}}(c)$
    end if
    Update $\{A_{b,h}\}$ using $\nabla_{A_{b,h}} \mathcal{L}$ ▷ gradient step with lr η
  end for
end for
return $\{A_{b,h}\}$

images generated by Flux. This cross-modal alignment motivates the design of SafeRoPE: *jointly intervening* on both image and text branches yields more reliable suppression of unsafe semantics than manipulating RoPE on text alone.

7.3. Rotary Positional Embedding (RoPE) in FLUX

FLUX.1 employs RoPE [33] for every attention head, which is inserted after Q/K projections and before the attention operation, as highlighted in the red boxes of Figure 8-(b1, b2).

8. Algorithmic Details

In this section, we provide additional algorithmic details of SafeRoPE. Pseudocode for unsafe subspace construction and safety-critical head selection is provided in Algorithm 1, and the full training procedure with risk-aware rotations is given in Algorithm 2.

8.1. Latent Risk Score (LRS) Calculation

For each attention head, given a query vector $q \in \mathbb{R}^{d_h}$ (similar for key vector), the LRS measures the proportion of the vector's energy lying in the unsafe subspace.

- q is fully unsafe: meaning $q = Uc$ for some coefficient vector $c \in \mathbb{R}^r$, then $Pq = UU^\top Uc = Uc = q, \Rightarrow \|Pq\|_2^2 = \|q\|_2^2$. Thus, $\text{LRS}(q) = 1$.
- q is fully safe: meaning $U^\top q = 0$, then $Pq = UU^\top q = 0, \Rightarrow \text{LRS}(q) = 0$.

8.2. Subspace Rotation Design

SafeRoPE intervenes via an orthogonal operator $\mathcal{R} = U\exp(A)U^\top + (I - UU^\top)$. Preserving the safe complement $(I - UU^\top)$ is necessary for the following reasons:

- Unnecessary distortion: Unsafe semantics occupy only a low-rank subspace ($r \ll d_h = 128$). Rotating the full head space would dramatically increase parameters and risk corrupting benign features. Constraining rotations to U ensures precise, localized intervention.
- Semantic fidelity: Vectors aligned with U (unsafe) are rotated by $U\exp(A)U^\top$, while vectors orthogonal to U (safe) pass through unchanged. The safe complement explicitly guarantees that benign components remain unaltered.
- Strict orthogonality: Rotating only $U\exp(A)U^\top$ does not yield an orthogonal map:

$$(U\exp(A)U^\top)^\top (U\exp(A)U^\top) = UU^\top \neq I.$$

  Adding the untouched safe complement completes the orthogonal transformation.

9. Implementation Details

9.1. Optimization and Hyperparameters

The hyperparameters used for training SafeRoPE are summarized in Table 3. All experiments adopt AdamW with mixed-precision (bf16) and a fixed image resolution of 1024. Beyond these settings, we highlight several practical details that are essential for stable reproduction.

9.2. Unsafe Token Collection and Filtering

To estimate reliable low-rank unsafe subspaces, we collect 1,000 unsafe query/key activations per attention head. Be-

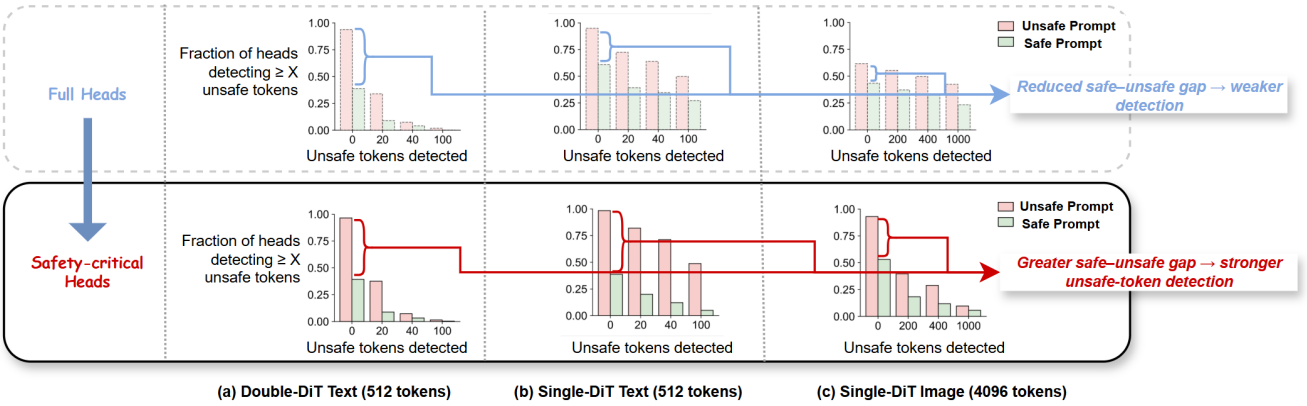


Figure 10. Comparison of unsafe-token detection across full heads and safety-critical heads. Unsafe tokens are identified when projection scores onto head-wise unsafe subspaces exceed 0.7. Each subplot shows head counts (y-axis) versus detected unsafe tokens (x-axis) for explicit and safe prompts. Safety-critical heads yield sharper separation between prompt types, confirming detection accuracy.

Table 3. Training Hyperparameters Used in SafeRoPE.

Hyperparameter	Value
Learning Rate	1×10^{-3}
LR Warmup Cycles	1
LR Scheduler Power	1
Optimizer	AdamW
Adam β_1	0.9
Adam β_2	0.999
Max Sequence Length	512
Mixed Precision	bf16

cause unsafe prompts vary in their ability to elicit meaningful activations, we follow a multi-stage filtering pipeline:

- Unsafe subject construction: Candidate subjects are drawn from public prompt repositories [†] and expanded using GPT-4o. We retain only subjects whose SBERT [56] similarity to a curated explicit seed set exceeds 0.5, ensuring semantic relevance.
- Template and modifier synthesis: We generate 30 neutral, composition-diverse templates using GPT-4o (e.g., "a realistic depiction of {} looking at the viewer"), designed so that unsafe characteristics arise solely from the inserted subject. Unsafe modifiers are taken from jailbreak-effective modifier sets in prior work [10] (e.g., "hyper-realistic, cinematic, NSFW, abstract art, horror art"), increasing the likelihood of triggering unsafe activations.

[colback=gray!5!white, colframe=gray!75!black, title=**Example of Unsafe Prompt Synthesis**, arc=2mm, boxrule=0.5pt] **Template:** A realistic depiction of {*subject*} looking at the viewer.
Subject (S)**:** a nude female figure
Modifier (M)**:** hyper-realistic, cinematic, NSFW, abstract art, horror art **Synthesized Prompt** ($\mathcal{P}$)**:** A realistic depiction of *a nude female figure* looking at the viewer, *hyper-realistic, cinematic, NSFW, abstract art, horror art*.

[†]https://huggingface.co/datasets/jtatman/stable-diffusion-prompts-stats-full-uncensored

9.3. Experiment Setup and Metrics

All experiments are conducted on FLUX.1-dev[‡] and FLUX.1-sch[§], which provide stable MMDiT implementation. SafeRoPE is model-agnostic and readily applicable to other FLUX variants (e.g., pro, kontext) due to their shared DiT-based attention structure.

Images are generated at 1024×1024 resolution using 28 rectified-flow sampling steps, without classifier-free guidance unless specified. We employ two established metrics to evaluate performance comprehensively.

- NudeNet Detation Rate: To quantify whether a generated image contains explicit nude content, we apply the NudeNet detector[¶], which outputs a set of confidence scores $\{s_k\}_{k=1}^{K}$ corresponding to predefined explicit categories (e.g., breast, buttocks, genitals, etc.). An image I is marked as unsafe (detected) if $\max_k s_k > 0.65$. The detection rate over an evaluation set $\mathcal{D}$ is then

$$\text{NudeNetRate} = \frac{1}{|\mathcal{D}|} \sum_{I \in \mathcal{D}} \mathbf{1}\left[\max_k s_k(I) > 0.65\right]$$

- CLIP Score: We compute the score using CLIP (ViT-L/32)[‖] to measure semantic alignment between text and image embeddings via cosine similarity. As a widely adopted metric in diffusion model evaluation [3], it effectively indicates whether unsafe semantics are suppressed while safe content is preserved.

[‡]https://huggingface.co/black-forest-labs/FLUX.1-dev
[§]https://huggingface.co/black-forest-labs/FLUX.1-sch
[¶]https://github.com/notAI-tech/NudeNet
[‖]https://huggingface.co/openai/clip-vit-base-patch32

- VQA Score: VQA Score is obtained with CLIP-FlanT5-XL[**], which assesses fine-grained semantic consistency, including object presence, spatial relationships, and structural integrity. This metric is particularly sensitive to unintended distortions caused by safety interventions, making it crucial for verifying that SafeRoPE maintains benign content fidelity.
- FID Score: We measure the overall visual quality and distributional realism using the Fréchet Inception Distance (FID)[††]. By computing the distance between the feature representations of generated and reference images extracted via a pre-trained Inception-v3 network, this metric is highly sensitive to low-level visual artifacts and diversity degradation. A lower FID score effectively validates that the concept unlearning process preserves the foundational generative capabilities of the model without compromising overall image fidelity.

10. Additional Experiments

10.1. Safety-Critical Heads Effectiveness

Safety-critical heads are identified by analyzing which heads yield consistent high-risk activations when text or image tokens are projected onto their unsafe subspaces. We label a token as unsafe when its projection score exceeds 0.7. As shown in Figure 10, filtering based on this criterion substantially increases the proportion of correctly identified heads across Double-DiT text, Single-DiT text, and Single-DiT image branches. These heads exhibit clear separation between safe and explicit prompts, validating their role as primary carriers of unsafe semantics and supporting head-wise targeted intervention in SafeRoPE.

10.2. Sensitivity to Positional Encoding.

We examine the model's sensitivity to positional encoding by applying random perturbations independently to text and image positional IDs. This analysis reveals how RoPE affects spatial–semantic fusion and generation quality under controlled positional disturbances. As shown in Figure 11, perturbing text positional IDs barely affects image fidelity but slightly disrupts long-text generation in generated image, implying that the T5 encoder already encodes basic positional information. In contrast, perturbing image positional IDs causes severe quality degradation, confirming FLUX's strong spatial dependence on RoPE. Further, random perturbations to RoPE text IDs under explicit prompts significantly suppress unsafe content while preserving visual fidelity for simple generations (Figure 12), leveraging RoPE's positional decay to weaken unsafe token coupling. However, such perturbations struggle with complex generations (e.g., long text) due to disrupted long-range dependencies. Compared to EraseAnything, which often fails under modifier-augmented unsafe prompts, RoPE perturbation provides a lightweight defense by blocking unsafe token co-activation without harming prompt semantics in simple cases.

[**]https://huggingface.co/zhiqiulin/clip-flant5-xl
[††]https://github.com/mseitzer/pytorch-fid

Figure 11. Effect of positional ID perturbations on Flux generation

Figure 12. Effectiveness of RoPE text position ID perturbations for safety alignment.

10.3. Trade-off Between Safety and Fidelity

We evaluate the joint effect of low-rank rotation and LRS, comparing SafeRoPE with EraseAnything (EA). Generated images are evaluated by NudeNet using: (i) *Non-hard*: any exposed-body label counted unsafe; (ii) *Hard*: only explicit labels[‡‡] counted. As shown in Table 4, For non-person classes, it matches or surpasses EA and FLUX.1-dev in CLIP and VQA metrics. On unsafe datasets, SafeRoPE achieves the lowest jailbreak rates in Table 4, despite no I2P-specific training. This demonstrates strong robustness and cross-dataset generalization.

10.4. Additional Qualitative Visualizations

We present extended qualitative comparisons among baselines in Figure 13, including SLD [58], UCE [14], DES [15], ESD [13], EraseAnything [18] and SafeRoPE on explicit prompts (Unsafe-1K, I2P [36]). SafeRoPE consistently preserves structural coherence and prompt semantics while suppressing only harmful content. In contrast,

[‡‡]*FEMALE-BREAST-EXPOSED*, *FEMALE-GENITALIA-EXPOSED*, *MALE-BREAST-EXPOSED*, *MALE-GENITALIA-EXPOSED*, *BUTTOCKS-EXPOSED*, *ANUS-EXPOSED*

	CLIP ↑				VQA ↑				Unsafe-1k ↓		I2P ↓	
Method	food	scenery	person	other	food	scenery	person	other	non-hard [%]	hard [%]	non-hard [%]	hard [%]
ERASEANYTHING	30.9	31.1	31.2	30.5	85.8	85.8	89.7	86.7	55.5	21.5	19.7	7.5
SafeRoPE	**31.1**	**31.4**	31.1	**30.6**	**89.8**	**88.1**	87.7	**87.0**	**36.7**	**15.4**	**12.5**	**7.0**
FLUX.1-DEV	31.4	31.5	31.3	30.8	87.2	86.0	88.4	85.9	73.7	38.8	25.2	10.3

Table 4. Quantitative comparison on category-specific alignment (CLIP, VQA; higher is better) and safety (Unsafe-1k/I2P; lower is better). Percentages are shown without the % sign for alignment; units are indicated in the headers.

FLUX.1-dev exhibits clear safety failures, and EraseAnything frequently introduces global artifacts or removes benign details. These qualitative results corroborate our quantitative findings and highlight SafeRoPE's ability to perform accurate, concept-localized safety intervention with minimal impact on non-target content.

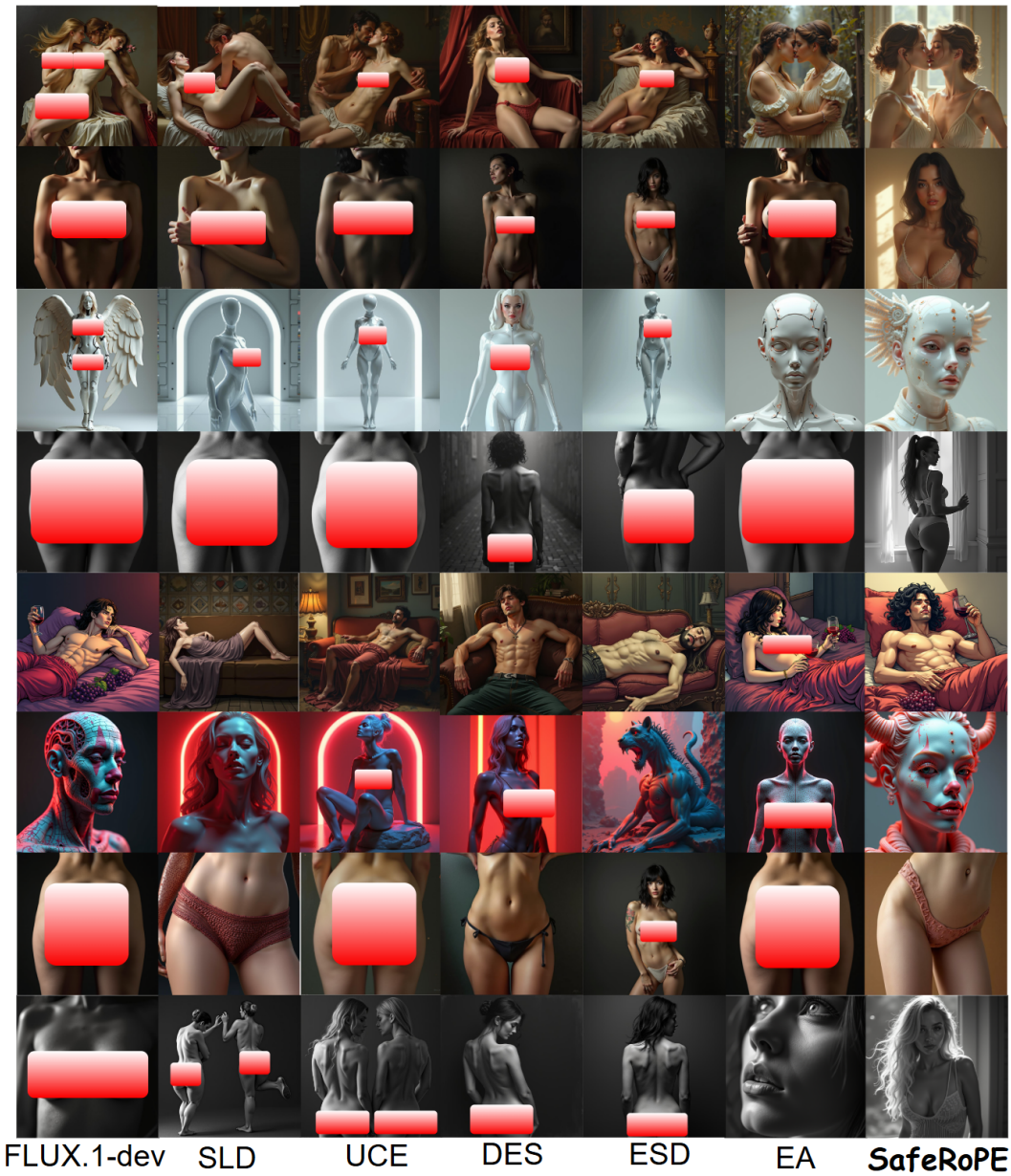


Figure 13. Qualitative comparison on unsafe prompts from Unsafe-1K and I2P..